CS224N Project Ideas

28 January 2011

Relation Extraction in the Knowledge Base Population (KBP) context
contact: Mihai Surdeanu <surdeanu@gmail.com>
contact: David McClosky <mcclosky@stanford.edu>

Here the task is to extract a set of "slots" for a given entity (person or organization), such as "schools attended", "date of birth", "important employees" etc. We already have a model for this, which we submitted at the shared task evaluation in 2010. So the students would not start from scratch. One important issue that is not resolved yet is how to address redundancy (and conflicts due to redundancy). For example, our system may extract "1880" as the date of birth for the entity "Pablo Picasso" once, and once mark this date as no relation in a different sentence. Which extraction should we trust in this common scenario?

The job of the students would be to evaluate at least three different approaches for this:
1. The mention-level model: this model handles each extraction separately during classification, and then unifies the outputs using an average (possibly weighted by classifier confidence).
2. The relation-level model: this model combines all mention feature vectors into a single vector (the relation vector). Hence redundancy for same slot value is avoided, because all mentions are collapsed into a single vector.
3. Last but most importantly, the URNS model proposed by the TextRunner group at UWash. (paper attached).
4. I would give big bonus points to any novel approach that beats all previous three approaches.

BioNLP Event Extraction
contact: David McClosky <mcclosky@stanford.edu>
contact: Mihai Surdeanu <surdeanu@gmail.com>

This project would be similar to performing named entity recognition for biomedical entities. However, instead of labeling entities with a small number of classes (Person, Organization, etc.), entities would be mapped to nodes in biomedical ontologies (a hierarchy of types). Given labeled training data, the model can be trained to take advantage of the structure of the ontology. For example, it may be able to learn which features can be shared across sibling ontology nodes. One baseline model is to use an NER system with only the leaf nodes in the ontology (essentially, ignoring all the hierarchical structure). A simple model would recursively classify each entity starting at the root node in the ontology. For example, the first classification decision would choose a child of the root node, the second classification decision would choose a child of the node selected in the first step, etc. until a leaf node is selected or a stopping decision has otherwise been made.

Dataset:

http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Technical+Term+Annotation

Ontology: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=GENIA+Ontology

Relevant literature: Jenny Rose Finkel and Christopher D. Manning. 2009. Nested Named Entity Recognition. In *Proceedings of EMNLP 2009*. [pdf]

A little biomedical domain knowledge would be helpful for this project, but likely not essential.

Predicting U.S. Elections with Twitter
contact: Nate Chambers <natec@stanford.edu>

Twitter provides instant access to the sentiment of millions of people, but determining the overall mood on any particular topic is an unsolved problem. Professional organizations rely on expensive human polling to track political trends, but it is unknown whether this can be automated through social media like Twitter. This project proposes using 6 months of Twitter data from 2009 to predict measures from human polls such as the Presidential Approval rating and the Generic Ballot (overall political leaning). You will analyze and model the language used in tweets to discover sentiment and political mood.

Various Projects Related to Litigation Analysis
contact: Ramesh Nallapati <nmramesh@cs.stanford.edu>

I. Outcome Prediction in Patent Litigation: The task is predict outcome of a case based on (i) past historical win rates of participants and (ii) merits of the case in terms of the strength of the patent infringed, etc. and (iii) ongoing trends in the case in terms of which party is able to win motions, etc. We have already done part (i) using a CRF to jointly model simultaneous cases involving the same party, and submitted this work to International Conference on AI and Law. Parts (ii) and (iii) need to be done, and will present interesting research challenges. Another interesting challenge is to use the cases that were settled as additional unlabeled data, since although they were settled, they usually have an implicit polarization towards plaintiffs or the defendants, which could be detected from individual motion outcomes.

II. Field Classification: This task is to classify cases into the technological field that the litigation is about. There are several types of inputs one can look into such as the court order documents, text of the asserted patents and the names of the parties involved. We already built a simple statistical classifier using all three features using a simple multi-task learning model that treats each input type as a task, and we have achieved about 65% F1 on the field of Mobile technologies. We also have gold data for SemiConductor and LCD fields, which we could use to build a transfer learning model across fields as well.

III. Early warning system for litigation: Analyze the patterns of litigation in the recent past and warn the companies that are actively engaged in the technology area of the litigation activity about the potential risk. May need both graph and textual analysis models.

IV. Attorney recommendation system: Suggest attorneys to parties involved in litigation based on technology field, opposing party, venue, and judge, and types of motions to be filed. Could use the hired attorneys in historical cases as the ground truth.

V. Product extraction: Extraction of products that infringe the asserted patents from court documents. We want to treat it as a weakly supervised problem with only a few seed examples and then bootstrap in an unsupervised manner. The novel legal domain may also pose additional research challenges.

VI. Entity Resolution: We have several types of entities such as parties, attorneys, law firms and judges. The task is to normalize the mentions of these entities in cases into their unique IDs. We already have a rule-based system in place that works well using name-normalization, but there is a lot of co-occurrence data (such as parties and attorneys tend to work together in several cases, etc), that could be exploited using a statistical model to boost performance.

Document classification to identify outbreak-related web content

Contact: Robert Munro <rmunro@stanford.edu>

At Global Viral Forecasting (www.gvfi.org) we are building a very large system to model and predict every contagion. Imagine Google Flu Trends, but with the goal of creating structured reports about all contagions and instead of looking at just search-terms you have all the world's available medical reports, news articles, blogs, social networks, citizen reports, field reports, etc.

An important step is identifying online content that relates to outbreaks. This project would be to investigate different methods for filtering relevant from irrelevant web-based documents. It will be a supervised learning task: we will give you a large collection of documents that are labeled as 'outbreak-related' or 'not outbreak-related' and you would build a system to that attempts to correctly identify outbreak-related documents among unseen test documents. There are number of potential directions for interesting research:

- what contribution do different features make to accuracy?
- what is the trade-off in accuracy/efficiency of different training algorithms (especially as this system will continually learn from new data)?
- what is the relationship between precision, recall and classification confidence?
- can you take advantage of any structured data in the documents (xml fields if RSS, HTML tags etc)?
- what is the loss in accuracy (if any) when trying to identify documents related to diseases that are not present in the training data?
- what is the relative accuracy across types of documents (eg: news articles vs medical reports)?
- to what extent should documents be treated differently according to the language of the report?

You could concentrate on any one or subset of these, or some other aspect that you thought was interesting.